DOCUMENT RESUME

ED 050 142 TM 000 529

AUTHOR Gleser, Leon Jay
TITLE The Attenuation Paradox and Internal Consistency.
INSTITUTION Johns Hopkins Univ., Baltimore, Md.
PUB DATE Feb 71
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York, February 1971

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS Factor Analysis, Item Analysis, *Mathematical Models, Measurement Techniques, *Mental Tests, Scoring, Statistical Analysis, *Test Construction, *Test Reliability, *Test Validity, True Scores
IDENTIFIERS *Attention Theory

ABSTRACT

An attempt is made to indicate why the concept of "true score" naturally leads to the belief that test validity must increase with an increase in test and/or average item reliability, and why this is correct for the classical single-factor model first introduced by Spearman. The statistical model used by Loevinger is introduced to establish the "attenuation paradox", and, in intuitive terms, attempt to explain why the "attenuation paradox" holds in this particular model. This is accomplished by showing that high (internal) consistency or reliability of test scores is an asset in increasing test validity under the classical single-factor statistical model for mental tests, but can be a liability when item scores are modelled as in the statistical model discussed by Loevinger. It is hoped that by this exposition, mental test specialists will be led to more critical appraisal of commonly used techniques and concepts (including the "corrections for attenuation"), and will check that their methods of test construction and comparison are consistent with their statistical models. (Author/PR)

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

ED050142

The Attenuation Paradox and Internal Consistency

by

Leon Jay Gleser [1/]

Department of Statistics
The Johns Hopkins University

The concept of "true score" lies at the heart of much of classical mental test theory and, as mentioned in the previous paper in this session (Finucci (1971)), is the basis of the derivation of "attenuation theory" (formulas which correct correlation coefficients for perturbing effects of errors of measurement). So much a part of the thinking of mental test specialists has the concept of "true score" become that the intuitions and consequences that can be derived from such a concept are frequently applied in situations where neither the "true score plus error" model nor the conclusions resulting from that model are applicable. In particular, misapplication of the "true score" concept seems to be behind the commonly held opinion that test validity can be increased by increasing test or item reliability. This opinion was shown by Loevinger (1954) to be false in a certain statistical model useful in item analysis of the dichotomously-scored items found in many aptitude tests. Loevinger (1954) named the assertion which she verified in her paper "the attenuation paradox".

1/ Presented at the American Educational Research Association 55th Annual Meeting, New York City, February 4-7, 1971.

TM 000 329

The Concise Oxford Dictionary defines "paradox" as a "statement contrary to received opinion ... seemingly absurd though perhaps really well-founded ..., conflicting with pre-conceived notions of what is reasonable or possible". Loevinger's (1954) "attenuation paradox" asserts that it is possible to 'attenuate" or reduce test validity with an increase in test and/or average item reliability. Although the word "attenuation" appears both in "attenuation paradox" and in "correction for attenuation", the connection between these concepts lies not so much in their use of a common word, but through the warning the "paradox" should give to practitioners who naively use the "correction for attenuation" formulas in inappropriate statistical contexts.

In the present paper, we first try to indicate why the concept of "true score" naturally leads to the belief that test validity must increase with an increase in test and/or average item reliability, and why for the classical single-factor model first introduced by Spearman (1904a) this belief is, in fact, correct. Next, we introduce the statistical model used by Loevinger (1954) to establish the "attenuation paradox", and in intuitive terms attempt to explain why the "attenuation paradox" holds in this particular model. We do this by showing that high (internal) consistency or reliability of test scores is an asset in increasing test validity under the classical single-factor statistical model for mental tests, but can be a liability when item scores are modelled as in the statistical model discussed by Loevinger. It is hoped that by this exposition, mental test specialists will be led to more critical appraisal of commonly used techniques and concepts (including the "corrections for attenuation"), and will check that their methods of test construction and comparison are consistent with their statistical models.

2. THE CLASSICAL MODEL

A central aim of mental test construction is to find a test which assesses with maximal accuracy the extent or level to which a given mental trait is possessed by an individual or individuals. In the classical statistical model of Spearman (1904a,b), it is assumed that the level of the mental trait in question can be measured by a single variable Y. Values of Y are assumed to have some probability distribution over that population of individuals which is of interest. Without essential loss of generality for this theoretical discussion, we may assume that Y has a mean (expectation) of zero and a variance of one.

If we could observe Y without error, there would, of course, be no need for a theory of mental tests (at least insofar as this theory refers to test construction). However, in general the trait level Y is not directly observable - it is latent. What we observe are scores $X_1, X_2, \ldots, X_N$ on N items. These items (sub-tests, questions, reaction times, etc.) are assumed to be statistically related to Y in that each item score individually can be used to predict or estimate Y by means of statistical regression techniques. For a given individual (given value of Y), it is assumed that the item scores $X_1, X_2, \ldots, X_N$ are (conditionally) statistically independent, and that given Y, the i^{th} item score X_i has (conditional) mean Y and (conditional) variance σ_i^2, $i = 1, 2, \ldots, N$.

The above assumptions relating the item scores $X_1, X_2, \ldots, X_N$, and Y are equivalent to a single-factor statistical model for the item scores, with Y as the common factor and each item score X_i having equal factor loading on Y. Consequently, we can assert that

$$X_i = Y + E_i, \qquad i = 1, 2, \ldots, N, \tag{1}$$

where Y, E_1, E_2,...,E_N are statistically independent, each E_i has mean equal to zero, and the variance of E_i is σ_i^2, $i = 1, 2, \ldots, N$. The model (1) is a "true score plus error" model for the item scores, and in this model, Y is the "true score".

To justify this model empirically it has been necessary for Spearman, Gulliksen (1950), and other mental test theorists to conceive of each item as being replicable on the same individual in such a way that the item scores X_i and X_i' on the i^{th} item and its replication are associated only through the fact that an individual brings the same mental trait level Y to bear on the replicated items. Stated statistically, these theorists have had to assume that an item could be paired with a supposedly parallel or identical item in such a way that the resulting item scores have the representation

$$X_i = Y + E_i,$$

(2)

$$X_i' = Y + E_i',$$

where Y, E_i, E_i' are independent, and E_i and E_i' have the same distribution. (Thus, E_i and E_i' both have mean zero and variance σ_i^2). Such assumptions are open to criticism, both in terms of the circularity in definiton required to operationally define parallel items (see Loevinger (1947, 1957), Ross and Lumsden (1968)), and in terms of difficulty of practical application (see Finucci (1971)). However, if accepted, these assumptions imply that if we could infinitely replicate an item, the average of the resulting item scores would equal Y. Hence, it seems that by maximizing internal consistency, we can almost perfectly estimate Y by choosing a test having a large enough collection of replicated items.

Item analysis aims at choosing items in such a way that maximal test validity is achieved with a minimal set of items. A mental test is thus a choice of items from a certain item pool of N items. Let τ denote the list of indices of items chosen (for example, τ might equal {1, 3, 9, 10, 12}). If the i^{th} item is used in our test, we write $i \in \tau$. The test score T is the sum of item scores over all items in the test; hence $T = \sum_{i\in\tau} X_i$. If there are n items, $n \leq N$, in the test, then

$$(3) \qquad \frac{1}{n} T = \frac{1}{n} \sum_{i\in\tau} X_i = Y + \frac{1}{n} \sum_{i\in\tau} E_i \equiv Y + E \quad .$$

Hence T/n also can be written in "true-score-plus-error" form. The "error" here is $E = \frac{1}{n} \sum_{i\in\tau} E_i$ which is statistically independent of Y, and has mean 0 and variance $\sigma^2 \equiv (\sum_{i\in\tau} \sigma_i^2/n^2)$.

To measure the accuracy with which the test score serves as an estimate of Y (i.e., the <u>validity</u> of the test), for theoretical purposes we may use the Pearson product-moment correlation coefficient ρ_{TY} between T and Y. Using formula (3) for T,

$$(4) \qquad \rho_{TY} = \rho_{(T/n)Y} = \frac{1}{(1 + \sigma^2)^{\frac{1}{2}}} \quad .$$

Since Y is unobservable, we cannot in practice estimate ρ_{TY} directly. Various sample measures of validity do exist (split-half validity, correlation with another test presumed to measure Y, etc.). However, these are fairly difficult to obtain in most cases. However, from formula (3) we see that T/n differs from Y by an error term E which is independent of Y and has mean 0 and variance σ^2. As σ^2 becomes smaller, E becomes less and less variable about its mean of 0. Thus justifies ρ_{TY} as a measure

of accuracy since $\rho_{TY} = 1/(1+\sigma^2)$ increases to 1 as σ^2 converges to 0 (see formula (4)). On the other hand, if we conceive of replicating each of the items in the test as in (2), we can think of a replicated test with test score $T' = \sum_{i \in T} X'_i = n(Y + E')$. We can thus measure the consistency, precision, or reliability of our test score by seeing how well T can predict its replicate T' (remember that both replicates are given to each individual). A measure of this predictability is the product-moment correlation $\rho_{TT'}$ which equals

$$\rho_{TT'} = \rho_{(T/n)(T'/n)} = \frac{1}{1+\sigma^2} \quad . \tag{5}$$

The fact that the reliability $\rho_{TT'}$ is inversely related to the variance σ^2 is intuitively obvious when we note that $T = T' + E - E'$. Since E and E' are independent, the variance of $E-E'$ is $2\sigma^2$. Hence the smaller the variation in the error term E (and its replicate E') is, the better able we are to predict T from T' (or T' from T). As σ^2 goes to 0, $\rho_{TT'}$ increases to 1 (see (5)), as is proper for a measure of reliability.

Comparing formulas (4) and (5), we see that

$$\rho_{TY} = \sqrt{\rho_{TT'}} \quad . \tag{6}$$

Consequently, we have verified mathematically that in the classical Spearman single-factor model, test validity increases monotonically with test reliability. However, this direct tie between test validity and test reliability occurs because in the "true score plus error" model satisfied by the test score T, the error term doubles as both an indicator of how accurately T measures Y and as an indicator of how repeatable the test score T is when the test is replicated on the same individual.

Before leaving the classical model, we pause to point out that test construction is often done by choosing those n items from the item pool which have maximum item reliabilities $\rho_{X_iX_i'}$. This practice of judging a choice of items solely by item reliabilities, rather than also by consideration of the correlations $\rho_{X_iX_j}$ between items (as would be necessary in multiple regression), is also a consequence of the classical "true-score-plus-error" model (1). Indeed if we calculate $\rho_{X_iX_i'}$ and $\rho_{X_iX_j}$, we find that

$$(7) \qquad \rho_{X_iX_i'} = \frac{1}{1+\sigma_i^2}, \qquad \rho_{X_iX_j} = \frac{1}{\sqrt{1+\sigma_i^2}\sqrt{1+\sigma_j^2}},$$

so that $\rho_{X_iX_j} = \rho_{X_iX_i'}^{\frac{1}{2}} \rho_{X_jX_j'}^{\frac{1}{2}}$. The extremely tight correlational structure revealed by this last mathematical result is not surprising, of course, when we recall that our model is a single-factor model. Remembering that $n^2\sigma^2 = \sum_{i\in T} \sigma_i^2$, and making use of (5) and (7), we find that the test reliability $\rho_{TT'}$ is a function solely of the item reliabilities; namely,

$$(8) \qquad \rho_{TT'} = \frac{n}{n-1 + \frac{1}{n}\sum_{i\in T}\left(\frac{1}{\rho_{X_iX_i'}}\right)}.$$

From formula (8), we see that if we want to choose the best test consisting of n items, we should choose the n items having highest reliability in our pool of items. Approximating the harmonic mean $\left[\sum_{i\in T}(1/\rho_{X_iX_i'})(1/n)\right]^{-1}$ by the arithmetic mean

$$(9) \qquad \bar{\rho} = \frac{1}{n} \sum_{i \in T} \rho_{X_i X_i'}$$

in (8), we only decrease $\rho_{TT'}$, but obtain

$$(10) \qquad \rho_{TT'} \geq \frac{n}{n-1 + \frac{1}{\bar{\rho}}} = \frac{n\bar{\rho}}{(n-1)\bar{\rho}+1} \quad .$$

The right side of this inequality is of course, the Spearman-Brown prophecy formula (see Finucci (1971)) commonly used for assessing test reliability.

3. THE NORMAL OGIVE MODEL

The classical statistical model described in Section 2 implicitly assumes that item scores are continuous variables. There is nothing in the model outlined in Section 2 to make this assumption necessary. Mental test tradition, however, has assumed that the mental trait level Y is a continuous random variable. Indeed, tradition further assumes that Y has a normal distribution. Despite challenges to this tradition (see, for example, Humphreys (1956)), most mental test theorists continue to adhere to the view that mental trait levels are continuously (normally) distributed. If this view is accepted, then a result from probability theory tells us that the representation (1) for item scores X_i implies that X_i must be a continuous random variable.

For most of the types of data originally considered by Spearman, item (or sub-test) scores were continuous variables (or could be thought of as rounded-off continuous variables). However, the basic items of modern

mental aptitude tests are multiple choice questions. These questions are customarily scored on a pass-fail, dichotomous basis ($X_i = 1$ or 0; or if we correct for guessing, $X_i = 1$ or $-1/k$, where k is the number of choices) Such item scores are clearly not continuous, except to a ridiculously gross approximation. Hence, we must either drop the "true-score-plus-error" model (1), or change our assumptions about the continuity of Y.

One attempt to preserve the basic features of the Spearman model, and yet retain the assumption that Y is normally distributed, is the normal ogive model. Here, we assume that to answer the i^{th} item in an N-item pool of dichotomously scored items, an individual calls upon a certain level of aptitude X_i which is available to him at that point for answering the item. This aptitude is assumed to be related to the level Y of the underlying mental trait of interest by a single factor model equivalent to the model in (1). However, it is additionally assumed that the level Y and the "error" E_i both are normally distributed variables. To pass the i^{th} item (obtain a score $S_i = 1$ on the item), the individual's level of aptitude must exceed a difficulty level a_i; otherwise, $S_i = 0$. Hence,

$$S_i = \begin{cases} 1 & \text{if } X_i \geq a_i, \\ 0 & \text{if } X_i < a_i. \end{cases} \tag{11}$$

Our interest still is to accurately measure Y for a given individual, but now the observables are $S_1, S_2, \ldots, S_N$ rather than $X_1, X_2, \ldots, X_N$.

If we plan to look for a "true-score-plus-error" model for the item scores, it soon becomes apparent that there is no way to write the i^{th} item score S_i in a true-score-plus-error form in such a way that the "true" term depends monotonically upon Y, the error term is independent of Y, and the two terms are statistically independent. For if such a representation

exists, the average of an infinite number of replications of the i^{th} item must equal the "true score". This limiting average, in the present model, is equal to the conditional probability that the i^{th} item is "passed" given $Y = y$, or

$$(12) \qquad P(S_i = 1 | Y=y) = 1 - \Phi\left(\frac{a_i - y}{\sigma_i}\right),$$

where $\Phi(z)$ is the probability that a standard $N(0,1)$ normal variable is less than or equal to z. The graph of this "true score" against y is an S-shaped curve called a "normal ogive" (which gives the model the name we have assigned to it), and by looking at this graph we see that the true score is indeed monotonic (but non-linear) in the mental trait level $Y=y$. Unfortunately, the "error" $S_i - P(S_i=1|Y=y)$ has (conditional) variance $[\Phi((a_i-y)/\sigma_i)][1-\Phi((a_i-y)/\sigma_i)]$ depending upon the "true score" (12), so that the "error" is not independent of the "true score". Hence, we must be prepared for consequences of the normal ogive model that seem "paradoxical" in terms of the "true-score-plus-error" model.

For example, under the normal ogive model there is an "attenuation paradox". To demonstrate this fact, we first point out that items have maximum reliability when their difficulty level is zero - that is, when they have .50 probability of being "passed". This assertion is true regardless of what correlation the required aptitude level X, has with the underlying mental trait level Y (see Sitgreaves (1961), Tucker (1946)). Hence, if we forget that we are dealing with a statistical model in which the "true-score-plus-error" model is not appropriate, and naively apply the results described in the previous section, we would decide to set the item difficulties of all of our items at 0. This would indeed mean that the test score

(13) $$S = \sum_{i \in T} S_i$$

would have maximum reliability. Further, item reliabilities (as measured by the phi coefficient between S_i and its repetition S_i') are monotonically increasing with the "reliability" coefficient $\rho_{X_i X_i'}$ of the aptitude called upon to pass the i^{th} item (Sitgreaves (1961, p. 20)). Since the validity of the X_i's for measuring Y increases to 1 as the "aptitutde reliability" $\rho_{X_i X_i'}$ increases to 1, this leads us to expect that if the average of the "aptitude reliability" coefficients,

(14) $$\bar{\rho}_{XX'} = \frac{1}{n} \sum_{i \in T} \rho_{X_i X_i'}$$

increases to one, so will the test validity ρ_{SY}.

Unfortunately (Tucker (1946), Loevinger (1954), Sitgreaves (1961)), this conclusion is false. Instead, as the average "aptitude reliability" $\bar{\rho}_{XX'}$ increases from 0 to 1, the test validity coefficient ρ_{SY} at first rises, then reaches a maximum, and then drops (attenuates) as the average "aptitude reliability" $\bar{\rho}_{XX'}$ continues to increase.

The following intuitive explanation for the phenomenon may give insight into the differences between the normal ogive model and the classical model. First note (see Gulliksen (1945)), that when the average "aptitude reliability" $\bar{\rho}_{XX'}$ equals one and all of the item difficulties are zero, then all of the item scores are 1 if Y is non-negative, and all of the item scores are 0 if Y is negative. In this case, all of the aptitudes X_i perfectly measure Y, the item scores are perfectly reliable and accurate measures, but what is actually measured by the item scores is merely the answer to the question: "Is Y non-negative?" Here, 100 items provide no

more information about Y than does one item, and the information provided does no more than identify the sign (plus or minus) attached to the magnitude of Y.

On the other hand, if we permit ourselves to use items which are less than perfectly reliable, and in fact assign item-difficulties of the form:

$$
\begin{aligned}
a_1 &= -3, \\
a_2 &= -3 + \frac{2}{33}, \\
a_3 &= -3 + \frac{4}{33}, \\
&\vdots \\
a_i &= -3 + 2\left(\frac{i-1}{33}\right), \\
&\vdots \\
a_{100} &= -3 + 6 = 3,
\end{aligned}
\tag{15}
$$

then when every aptitude λ_i is exactly equal to Y, i = 1,2,...,100, a test score of S = k, $1 \leq k < 100$, tells us that the first k item scores $S_1, S_2, \ldots, S_k$ are all 1 and that the last 100-k item scores $S_{k+1}, S_{k+2}, \ldots, S_{100}$ are all 0. Why? Because $S_i = 1$ if and only if $Y \geq a_i$, and since $a_1 \leq a_2 \leq \cdots \leq a_{100}$, the fact that Y is greater than or equal to a_j ($S_j = 1$) implies that $Y \geq a_i$ for all $i \leq j$ ($S_i = 1$, all $i \leq j$), whereas if $Y < a_\ell$ for some ℓ, then $Y < a_i$ for all $i \geq \ell$. In other words, S and Y are monotonically related. Further, if we know that S = k, $1 < k < 100$, we can show that $-3 + (2(k-1)/33) \leq Y < -3 + (2k/33)$, while S = 0 means that $Y < -3$, and S = 100 means that $Y \geq 3$. Clearly these 100 items, although each is less reliable than the items whose difficulties are all zero, provide greater validity for measuring Y.

From the above discussion, we see that if we know that the average "aptitude reliability" $\bar{\rho}_{XX'}$ is close to one, we are at a disadvantage in terms of test validity if we must keep item difficulties the same ($a_i = 0$, $i = 1,2,\ldots,100$), even though item reliability may be maximized. Hence, the maxims of Section 2 provide no guideline in this case. If we are required to set all item difficulties equal to zero, then some other mechanism is needed to provide information about Y similar to that provided by the spread-out choices (15) for the a_i's. Amazingly enough, and in contrast to our use of the word "error", when average "aptitutde reliability" is less than one, the "errors" $E_i = X_i - Y$ provide this mechanism and allow us to increase test validity. If we replace the word "error" by "randomization", this result should not surprise statisticians (who know that controlled randomization in sample survey and experimental design can improve accuracy of measurement), but it certainly will surprise anyone who is used to thinking of the error E_i in the "true-score-plus-error" model as a source of lack of consistency and inaccuracy for measurement of Y. Nevertheless, in the present model a certain amount of "error" helps improve validity. Remembering that $a_1 = a_2 = \ldots = a_{100} = 0$ and that $\bar{\rho}_{XX'} < 1$, let us in fact assume for convenience that $\rho_{X_i X_i'} \equiv \rho$, all i. Looking back at the definition of S_i (Equation (13)), we see that we can rewrite S_i in the form

$$
\begin{aligned}
S_i &= \begin{cases} 1 & \text{if } Y + E_i \geq 0, \\ 0 & \text{if } Y + E_i < 0, \end{cases} \\
&= \begin{cases} 1 & \text{if } Y \geq -E_i \\ 0 & \text{if } Y < -E_i, \end{cases}
\end{aligned} \tag{16}
$$

$i = 1,2,\ldots,100$. Since Y and E_i are independent, (16) shows that the E_i's act as a random allocation of item difficulties for a new normal ogive model

in which the item aptitudes X_i are exactly equal to Y. Since almost surely the values of $E_1, E_2, \ldots, E_{100}$ are unequal (and in fact are a sample from a normal distribution with mean 0 and variance $(1-\rho)/\rho$), it seems reasonable that there is some value of the common "aptitude reliability" ρ (or range of values for ρ) where this random choice of item difficulties will improve test validity over the fixed choice $a_i = 0$, all i. If ρ is near 0, the variance of each E_i is nearly infinite, and the random item difficulties E_i will be too far spread out to provide much information about Y. (This can also be seen by remembering that $X_i = Y + E_i$, and noting that when the variance of E_i is near infinity, Y is basically unobservable.) For ρ near 1, the variance of each E_i is nearly 0, and thus E_i varies only very little from its mean of 0, so that this case is essentially that of fixed item difficulties. Hence, the value of ρ that will allow us to improve upon the fixed item difficulty, perfect reliability case, lies somewhere between $\rho=0$ and $\rho = 1$. Mathematically it is found that for a 100-item test, maximum test validity of $\rho_{SY} = .9729$ occurs for an (average) "aptitude reliability" of $\rho = .2268$ (Sitgreaves (1961)).

The above discussion provides an example where an acceptance of "error" helps to improve accuracy of measurement. It also indicates that deviation from the classical "true-score-plus-error" model, no matter how seemingly trivial these deviations are, may have major consequences for the theory and interpretation of indices of test performance. In any testing problem, therefore, the mental test specialist would do better to base his methods and conclusions on the statistical model, rather than trusting to intuition obtained from the classical "true-score-plus-error" model to guide his thinking.

ACKNOWLEDGEMENT

The work on this article was sponsored in part by the Air Force Office of Scientific Research, Air Force Systems Command, USAF under AFOSR Contract F44620-70-C-0060. The United States Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright hereon.

REFERENCES

(1) Brogden, H. E. Variation in test validity with variation in the distribution of item difficulties, number of items, and degree of their intercorrelation. Psychometrika, 1946, 11, 197-214.

(2) Fimucci, Joan M. Early history and basic derivation of attenuation formulas. Paper presented at American Educational Research Association 55th Annual Meeting, New York City, February 5, 1971.

(3) Gulliksen, H. O. The relation of item difficulty and inter-item correlation to test variance and reliability. Psychometrika, 1945, 10, 79-91.

(4) Gulliksen, H. O. Theory of mental tests. New York: Wiley, 1950.

(5) Humphreys, Lloyd G. The normal curve and the attenuation paradox in test theory. Psychological Bulletin 1956, 53, 472-476.

(6) Loevinger, Jane. A systematic approach to the construction and evaluation of tests of ability. Psychol. Monogr., 1947, 61, No. 4 (Whole No. 285).

(7) Loevinger, Jane. The attenuation paradox in test theory. Psychological Bulletin 1954, 51, 493-504.

(8) Loevinger, Jane. Objective tests as instruments of psychological theory. Psychological Reports, 1957, 3, 635-695. Southern Universities Press 1957. Monograph Supplement 9.

(9) Ross, John and Lumsden, James. Attribute and reliability. British Journal of Mathematical and Statistical Psychology 1968, 21 part 2, 251-263.

(10) Sitgreaves, Rosedith. A statistical formulation of the attenuation paradox in test theory (Chapter 1, pp. 17-28). Optimal test design in a special testing situation (Chapter 2, pp. 29-45). Further contributions to the theory of test design (Chapter 3, pp. 46-63). In Studies in Item Analysis and Prediction (ed. Herbert Solomon). Stanford: Stanford University Press, 1961.

(11) Spearman, C. The proof and measurement of association between two things. American Journal of Psychology, 1904, 15, 72-101.

(12) Spearman, C. "General intelligence" objectively determined and measured. American Journal of Psychology, 1904, 15, 201-293.

(13) Tucker, L. R. Maximum validity of a test with equivalent items. Psychometrika 1946, 11, 1-13.